Journal on
Artificial Intelligence

Tech Science Press

https://doi.org/10.32604/jai.2026.077823

ARTICLE

Artificially Intelligent Interviewer—A Multimodal Approach

Daniil Kamakaev and Khaled Mahbub*

Birmingham City University, 4 Cardigan Street, Birmingham, UK

*Corresponding Author: Khaled Mahbub. Email: khaled.mahbub@bcu.ac.uk

Received: 17 December 2025; Accepted: 26 February 2026; Published: 15 April 2026

ABSTRACT: This paper presents an innovative system designed to automate the analysis of candidate interviews by integrating multiple analytical techniques into a single multimodal framework. This system combines text sentiment analysis, audio sentiment analysis, keyword extraction, and Mel-Frequency Cepstral Coefficients (MFCC) feature extraction to evaluate candidate performance holistically. This system employs text sentiment analysis using VADER and transformer-based sentiment features (probability-based outputs), audio sentiment analysis with an SVM model trained on both IEMOCAP and MELD datasets, keyword extraction via KeyBERT, and audio feature extraction including MFCCs, delta MFCCs, pitch, and energy to evaluate candidate performance holistically. A novel weighted scoring mechanism, incorporating personality traits such as Neuroticism, distinguishes this project from existing single-modality systems, offering a comprehensive candidate assessment. Results highlight its potential for Human Resources (HR) applications, with future improvements planned to enhance Negative sentiment detection through hybrid approaches and advanced emotion recognition techniques.

KEYWORDS: VADER (Valence Aware Dictionary and sEntiment Reasoner); MFCC (Mel-Frequency Cepstral Coefficients); KeyBERT; RoBERTa; openSMILE (Open-Source Media Interpretation by Large feature-space Extraction); MELD (Multimodal EmotionLines Dataset); CMU-MOSI (CMU Multimodal Opinion Sentiment Intensity); IEMOCAP (Interactive EMOtional dyadic motion CAPture database)

1 Introduction

Interviews are central to organizational recruitment systems, assessing candidates' skills, personality, and cultural fit [1]. Evaluating candidate performance during interviews is a critical yet challenging task for Human Resources (HR) professionals, often relying on subjective judgments of emotional tone, content relevance, and vocal delivery. Subjective human judgment often introduces bias, necessitating AI/ML-based systems for objective evaluation [2]. AI/ML-based tools are gradually becoming popular among organizations for recruitment, as these tools may facilitate fair and consistent recruitment at reduced time and cost per hire [3–6]. AI/ML-based tools support human interviewers at different stages of the recruitment process by providing analytics and feedback. For example, at the initial stage of the recruitment process, these tools can support pre-planning, sourcing, resume screening, interview scheduling, etc., and AI/ML-based tools are now being increasingly used by organizations at this stage of the recruitment process [3,6].

Nevertheless, application of AI/ML during the interview stage of the recruitment process is relatively under-researched, despite AI/ML can scrutinize factors like word choice, tone to perform emotion and sentiment analysis, and can help interviewers avoid bias and maintain interview consistency [3,4]. Natural Language Processing (NLP) offers various transformer models like BERT [7] and DistilBERT [8] that

are capable of extracting meaningful information from text input, which is essential for sentiment and personality analysis in interviews. Similarly, tools like VADER (Valence Aware Dictionary and sEntiment Reasoner) can analyse text sentiment effectively [9]. However, these models and tools overlook audio features like tone or pitch; similarly, lexicon-based sentiment tools can struggle with context, sarcasm, and domain-specific phrasing; therefore, transformer-based sentiment models (e.g., RoBERTa) can be incorporated to capture richer contextual cues [7]. On the other hand, Speech Emotion Recognition (SER) is the process that deduces emotional states from vocal signals and is particularly valuable for evaluating delivery quality and authenticity in interview settings. Traditional SER methods typically depend on acoustic features that are handcrafted, and Mel-Frequency Cepstral Coefficients (MFCCs) are one such method that is most widely used due to their ability to capture speech spectral properties linked to emotional expression [10]. openSMILE, introduced in [11], is a versatile feature-extraction toolkit that provides MFCCs alongside additional descriptors such as pitch, energy, jitter, and shimmer, enabling a comprehensive characterization of vocal attributes relevant to emotion classification. These audio-focused systems, such as traditional speech recognition with MFCCs, often lack sentiment integration or are limited to speaker identification [12]. In literature, several commercial and academic solutions can be found that supports interview process, but most of these only rely on unimodal inputs or relatively shallow analysis. Among the popular existing automated interviewing systems, HireVue [13] and Mya Systems [14] can be mentioned here. HireVue uses video-based evaluations that draw on facial expressions and vocal features, but offer limited integration of textual data or fine-grained acoustic emotion signals. This lack of transparency raises interpretability and fairness concerns in automated hiring decisions. Mya Systems, in contrast, operates mainly as a conversational chatbot that gathers key data from job seekers through initial conversations with job seekers. This tool does not assist in actual candidate evaluation or sentiment analysis [15]. In summary, traditional automated interviewing tools typically focus on single modalities like text or audio and use fixed-rule approaches, miss the holistic perspective needed for HR, and lack flexibility to adapt to diverse requirements [13,16].

In this paper, we present the AI Interviewer, which focuses on developing an automated system to analyse candidate interviews by integrating text and audio sentiment analysis, personality assessment, and audio feature extraction. The methodology involved designing a multimodal framework using VADER for text sentiment analysis [9], an SVM (Support Vector Machine) model retrained on combined IEMOCAP (Interactive Emotional Dyadic Motion Capture) [17] and MELD (Multimodal EmotionLines Dataset) datasets for audio emotion detection [18], and KeyBERT for keyword extraction [19]. Audio features include MFCCs, delta MFCCs, pitch, and energy, extracted using librosa [20]. In addition, a transformer-based sentiment model is used to derive probabilistic sentiment cues that are incorporated into the sentiment classification stage. Unlike existing single-modality systems, in our framework, we employ a novel weighted scoring mechanism that incorporates Big Five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN), to deliver a more holistic candidate assessment. We implemented a GUI (Graphical User Interface) to facilitate user interaction, and evaluated the system on the MELD test dataset. Evaluation results demonstrate that our framework achieved improvement over a unimodal baseline system in terms of classification accuracy, which highlights its potential for Human Resources (HR) applications. The rationale for this work lies in addressing the lack of integrated multimodal tools in HR, aiming to reduce bias and enhance efficiency in candidate assessments through objective, data-driven analysis.

The rest of the paper is structured as follows, in Section 2 we provide an overview of recent developments in the area of automated interview analysis including application of Natural Language Processing (NLP) and Speech Emotion Recognition (SER) for interview analysis, in Section 3, we discuss the design and development of our modular framework for automated interview analysis, in Section 4 we present and

discuss the experimental results that were carried out to evaluate our framework and finally in Section 5 we provide concluding remarks and recommendations for future enhancements of our modular framework.

2 Related Work

The literature review explores the current state of research in automated interview analysis, focusing on natural language processing (NLP), speech emotion recognition (SER), and multimodal systems, and highlights gaps addressed by our approach, including a VADER + openSMILE baseline comparison and transformer-assisted sentiment features evaluated on MELD and CMU-MOSI [21].

2.1 NLP in Interview Analysis

Natural Language Processing (NLP) has revolutionised text analysis, providing tools to understand and extract meaningful information from interview transcripts with very high accuracy. Transformer models like BERT [7], DistilBERT [8] excel in capturing contextual nuances, with possible layered meanings, that are vital for sentiment and personality analysis in interviews. In [22] sentiment analysis techniques are discussed with focus on their effectiveness in customer reviews, which can be easily applied to interview settings with similar conversational dynamics. Recursive deep models for semantic compositionality are highlighted in [23]. These models prove useful for understanding complex interview dialogues where responses may carry multiple interpretations. VADER is a rule-based sentiment analysis tool optimised for conversational text introduced in [9]. We have adopted VADER in this project for its efficiency in handling short interview responses, ensuring quick processing that is vital for live HR applications.

In interview contexts, NLP enables detection of subtle linguistic cues, such as confidence markers, politeness strategies, and emotionally charged phrases that may play significant role in recruitment decision-making. NLP may also help to deduce behavioural profile and assess candidate suitability for specific roles. For example, linguistic features like adjective density, modality use, and discourse cohesion may contribute to identify traits such as assertiveness or hesitation, which are essential ingredients of behavioural profile.

While many NLP models offer high accuracy on benchmark datasets, few are tailored for real-time use in recruitment settings, where speed is paramount. The AI Interviewer bridges this gap by balancing performance and efficiency, prioritising practical deployment. For example, VADER was chosen over more complex transformer-based models due to its low latency and strong performance on short, informal utterances that are typical in live interviews. This ensures HR professionals receive timely insights, while more context-aware transformer approaches can be considered when higher latency is acceptable.

2.2 Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) identifies emotional states from vocal signals, which can be used during interviews to assess delivery and authenticity. Mel-Frequency Cepstral Coefficients (MFCCs) are crucial, traditional audio features for SER, which are widely used for capturing spectral characteristics of speech that correlate with emotional tone [10]. openSMILE introduced in [11] is a comprehensive toolkit that facilitates feature extraction, including MFCCs, pitch, energy, jitter, and shimmer, offering a rich representation of vocal attributes relevant to emotion classification.

Recent research trends highlight the efficacy of deep learning models in the SER process. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and the advanced version of RNN, i.e., Long Short-Term Memory (LSTM) architectures, have been shown to outperform traditional classifiers by capturing temporal dependencies and higher-level abstractions in vocal expression. Attention mechanisms,

further enhance performance, particularly in noisy or real-world environments by focusing on emotionally salient segments of audio [24].

In our approach we have used an extended version of MFCCs with additional temporal and prosodic features. Specifically, in our approach delta and delta-delta MFCCs were included to capture variations in speech dynamics over time [12,25]. Pitch and energy features were extracted to reflect intonation and vocal strength, which are critical indicators of emotional state. The model was implemented using a Support Vector Machine (SVM), which was trained on a combined corpus from the IEMOCAP and MELD datasets to improve generalisability and address potential domain mismatch issues between scripted and spontaneous emotional speech.

One challenge in SER is the presence of inter-speaker variability and background noise, which can obscure emotional cues. In this project we addressed such variability through audio pre-processing that includes noise reduction via spectral subtraction and standardisation to a uniform sampling rate (16 kHz). Furthermore, integrating textual sentiment and emotion features alongside audio signals created a more robust, multimodal framework for emotion recognition, mitigating the limitations of unimodal SER approaches.

2.3 Multimodal Systems

Multimodal systems combine text, audio, and sometimes video for enhanced analysis. This allows us to leverage the complementary strengths of each modality to achieve a deeper understanding of complex data. Superiority of multimodal fusion over unimodal approaches in affective computing have been studied and proven in the literature [26]. Recent studies in multimodal affective computing and sentiment modelling also highlight the role of transfer learning and temporal modelling in improving robustness across settings. In [27] analysis of YouTube movie reviews has been presented, where audio-visual integration improved sentiment classification. Late fusion and hybrid methods, as explored in [28], enhance emotion detection by integrating features at different stages, ensuring that diverse emotional cues are captured effectively. In [29] multi-tier deep learning framework is used for multimodal sentiment analysis, where CNN applied for extracting text-based features, 3D-CNN model for extracting visual features and openSMILE tool kit applied for audio feature extraction. This approach provides robust performance but its applicability to real time scenario is questionable as it requires intensive computational power. Similarly, in [30] a transformer-based model is presented that extends traditional transformer architectures by incorporating linguistic, cultural, and code-mixing attributes to detect context and cultural idioms. However, this approach only focuses on Tamil and Malayalam languages.

Inspired by this, in our work we fuse text sentiment, audio sentiment, and audio features like pitch and energy, aiming to address the need for integrated interview analysis in HR contexts. As discussed in Section 3, the weighted scoring system and personality-based adjustments are treated as configurable parameters to support different HR requirements, rather than being presented as universally optimal. This aligns with late fusion techniques whilst adapting to individual candidate traits. Our approach not only improves accuracy but also tailors evaluations to reflect nuanced emotional and behavioural patterns, which provides a foundation for more personalised HR assessments.

2.4 Existing Tools and Their Limitations

Several commercial and academic tools can be found in the literature that have been developed for candidate assessment. Nevertheless, most of these tools focus on unimodal or shallow analysis. Platforms like HireVue [13] and Mya Systems [14] are examples of this trend. HireVue employs video-based assessments using facial and tonal cues. However, it often operates as a black-box model, with limited integration of

textual content or detailed acoustic emotion cues. This opacity raises concerns about interpretability and fairness in automated decision-making processes. Mya Systems, on the other hand, primarily functions as a conversational AI assistant for recruitment scheduling, and it offers minimal involvement in candidate evaluation or sentiment detection [15].

These limitations highlight that unimodal and rule-based systems are inadequate to accurately interpret nuanced interview responses, as in interview process tone, context, and emotional regulation are essential. Moreover, tools that lack integration of modalities (e.g., text, audio, visual) are prone to produce biased or incomplete evaluations, as these tools may fail to capture the interplay between what a candidate says and how it is delivered.

The system developed in this project addresses these issues through a multimodal architecture that fuses textual sentiment, acoustic features (MFCCs, pitch, energy), and personality trait estimation to provide a richer and more reliable evaluation. By bridging the analytical gap left by existing tools, the AI Interviewer improves on both performance metrics and interpretability, offering a more transparent and ethically defensible solution for automated candidate assessment.

3 Design and Implementation of AI Interviewer

In this section we discuss the design and implementation of AI Interviewer. We follow a structured process to implement the framework, which is shown in Fig. 1. As shown in the figure, the structured process integrates multiple stages, starting from data input to model evaluation and scoring.

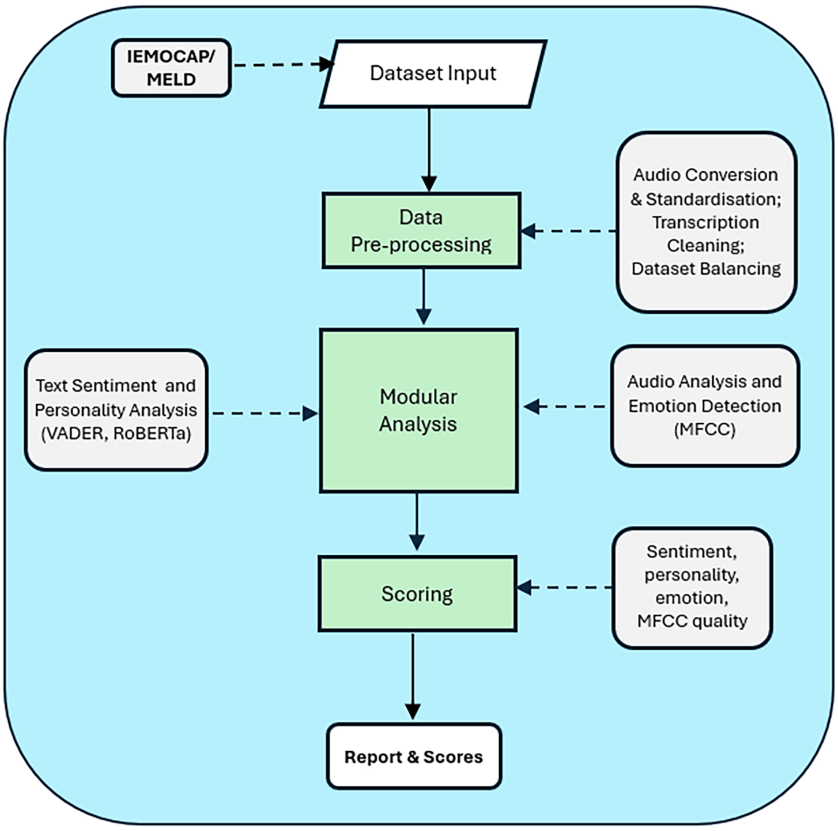

Figure 1: Process followed to implement artificially intelligent interviewer.

The AI Interviewer is designed and developed as a multimodal evaluation framework that integrates Speech Emotion Recognition (SER) and Natural Language Processing (NLP) to automate the assessment of spoken interviews. It recognises that human communication encompasses both verbal and non-verbal cues. Hence, our framework processes WAV audio files and corresponding transcripts to evaluate content, tone, and emotional delivery, and it supports both rule-based and transformer-assisted sentiment signals as part of the overall analysis pipeline.

In our framework VADER has been used to perform text sentiment analysis. We have selected VADER for its efficiency to analyse short conversational responses [9]. In addition, a transformer-based sentiment classifier (RoBERTa) was evaluated and used to generate probability-based sentiment features (e.g., class probabilities and confidence margins) that complement VADER in capturing deeper contextual patterns. Transformer-based models like DistilBERT [8] was another candidate for text sentiment analysis, but it was not used in the implementation due to computational constraints [7].

Audio sentiment is derived from Mel-Frequency Cepstral Coefficients (MFCCs), delta MFCCs, pitch, and energy (e.g., extracted via openSMILE), processed by a Support Vector Machine (SVM). The SVM emotion classifier was trained on a balanced combination of IEMOCAP and MELD datasets to improve robustness across scripted and conversational speech styles [17,18]. CMU-MOSI [21] was used separately as an auxiliary benchmark for text-based sentiment evaluation. Big Five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN) [31], are estimated from sentiment trends, providing behavioural insights.

The system employs a weighted scoring mechanism comprising sentiment, personality, emotion, and MFCC quality as described in Section 3.4, with default weights that are configurable to support different HR requirements.

In the rest of this section, we discuss each step of the structured process in detail.

3.1 Data Preprocessing

In our implementation we used two datasets: (a) the IEMOCAP dataset [17] for training and (b) the MELD dataset [18] for both training and evaluation. The IEMOCAP dataset provides dialogues annotated by categorical emotions, which enhances the model's ability to detect nuanced emotions. The MELD dataset comprises dialogue utterances with transcriptions and corresponding MP4 audio files. Each utterance in the dataset has sentiment (Positive, Negative, or Neutral) annotation. In addition, CMU-MOSI was used as an auxiliary benchmark dataset to evaluate text sentiment approaches (VADER and a transformer-based classifier).

Preprocessing was critical to ensure data consistency and following actions were taken during data preprocessing:

- Audio Conversion: MP4 files from the MELD dataset were converted to WAV format to enable audio feature extraction.
- Standardisation: To ensure uniform processing, audio files were standardised to a 16 kHz sampling rate [10].
- Transcription Cleaning: To improve text analysis accuracy, transcriptions were cleaned to remove inconsistencies, such as eliminating non-verbal annotations (e.g., "[laughter]") and correcting punctuation.
- Dataset Balancing: The training dataset, combining IEMOCAP and MELD, was balanced by subsampling to ensure equal representation of emotions (happy, sad, angry, neutral). This was done to mitigate class imbalance issues that could skew model performance.

3.2 Modular Analysis—Text Sentiment and Personality Analysis

Sentiment analysis in the AI Interviewer system is conducted using VADER sentiment analyser. This rule-based model offers transparency in its output, making it suitable for real-time HR applications where explainability is critical. VADER computes a compound score, which is mapped to Positive (>0.05), Negative (<−0.05), or Neutral sentiments. VADER's ability to swiftly analyse brief utterances ensures timely feedback, which is essential for dynamic interview settings where HR professionals require immediate insights to guide decision-making processes. In addition, a transformer-based sentiment classifier (RoBERTa) was used to generate probability-based sentiment features (e.g., negative/neutral/positive class probabilities, confidence, and margin), which complement VADER when deeper contextual signals are present.

In our implementation we have also used KeyBERT [19] library that employs BERT embeddings to extract keywords and key-phrases from text. In AI Interviewer KeyBERT identifies the most relevant terms from each response while maintaining low computational overhead. The extracted keywords are included in the final PDF report to represent the context of the interview. HR teams can use these keywords to understand the focus of discussions and interpret the sentiment analysis in a better way.

Personality trait estimation is based on sentiment trends mapped to the Big Five personality traits—Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (OCEAN)—and is used as an indicative layer rather than a full personality assessment [31]. This should be noted that association between personality traits and performance depends on many factors including occupational groups, disciplines, performance criteria, interpersonal sensitivity, etc. [32–36]. We have been inspired by the analysis done in [32–34] and derived scores for personality traits from aggregated sentiment patterns using mappings as shown below:

- Openness: sentiment_score * 1.00
- Conscientiousness: sentiment_score * 1.10
- Extraversion: sentiment_score * 1.05
- Agreeableness: sentiment_score * 1.20
- Neuroticism: (1 − sentiment_score * 0.80).

As shown in the above mappings, the multipliers reflect broad tendencies associated with each trait. For example, Agreeableness is given a higher proportional value because it is commonly associated with cooperative and positively framed responses, while Extraversion receives a moderate increase to reflect outward emotional expression. Neuroticism is modelled inversely, since variability or negativity in sentiment often aligns with higher emotional reactivity. It should be noted that we have used the term Personality Traits in the scoring formula, but these mappings do not aim to replicate full personality assessment but instead provide a simple behavioural tendency that complements the sentiment and audio analysis.

3.3 Modular Analysis—Audio Analysis and Emotion Detection

The audio analysis is used in AI Interviewer to enhance understanding of candidate responses by interpreting vocal cues. It extracts acoustic features (as described below) representing both the spectral and temporal properties of speech, which are crucial for identification of emotional states [10]. In our implementation, these features are extracted using an openSMILE-based pipeline to ensure consistent and reproducible acoustic descriptors across recordings:

- MFCCs (Mel-Frequency Cepstral Coefficients): This is used to identifying the emotion of the speaker from the speech signal [10,37,38].
- Delta MFCCs: Represent the rate of change of MFCCs, providing dynamic information about speech signals.

- Pitch: Fundamental frequency of the voice that indicates emotional states.
- Energy: Measures the loudness of speech that is used to analyse emotional intensity [39].

To ensure robust emotion classification, a Support Vector Machine (SVM) model was trained by using a carefully curated and balanced combination of the IEMOCAP and MELD training datasets. These datasets were combined to address domain mismatch, for example IEMOCAP provides a rich variety of well-labeled emotional utterances in both scripted and improvised settings [17], while MELD offers dialogue-level emotional context and conversational flow [18]. This dual-source training strategy improves generalisability across spontaneous and formal interview styles. Configuration used for the SER SVM in our experiments is summarised in Table 1 below.

Table 1: Speech Emotion Recognition (SER) SVM settings.

Setting	Value	Notes
estimator_class	SVC	Final estimator type
kernel	rbf	Kernel type
C	1.0	Regularisation strength (C)
gamma	Scale	Kernel coefficient (gamma)
class_weight	nan	Class weighting strategy
Probability	True	Whether probability estimates are enabled
decision_function_shape	ovr	Decision function shape (ovo/ovr)
tol	0.001	Stopping tolerance
max_iter	–1	Maximum iterations (–1 means no hard limit)
random_state	nan	Random seed (if set)

The selected SVM configuration reflects a deliberate balance between expressive capacity and stability for speech emotion recognition in conversational interview data. The RBF kernel was chosen due to its ability to model non-linear relationships in high-dimensional acoustic feature spaces, which is well-suited to MFCC-based representations. Hyperparameters were fixed rather than extensively tuned to prioritise reproducibility and avoid dataset-specific overfitting, given the cross-dataset training setup combining IEMOCAP and MELD. Probabilistic outputs were enabled to support downstream multimodal fusion, allowing emotion predictions to act as calibrated modifiers rather than hard labels within the final scoring mechanism.

The trained model classifies utterances into four primary emotional categories: happy, sad, angry, and neutral. These categories were derived from standard emotion taxonomies for their relevance in conversational contexts [40]. These emotion labels are used not only for standalone classification but also to support sentiment disambiguation, particularly in cases where verbal cues are ambiguous, enabling accurate interpretation of emotional intent. For example, neutral-sounding phrases delivered with a frustrated tone may be reclassified as negative when acoustic features suggest anger or distress.

In cases where text sentiment is ambiguous or borderline, the emotion prediction and Neuroticism scores act as modifiers. For example, a high Neuroticism score (>0.7) coupled with acoustic indicators of stress may shift a borderline Neutral sentiment towards Negative, improving sensitivity to subtle dissatisfaction or anxiety, ensuring the system accounts for emotional subtleties often missed in interviews. This adaptive mechanism enhances the system's ability to detect nuanced emotional states, providing a richer candidate profile for HR decision-making.

This integration of speech emotion recognition ensures that the AI Interviewer captures both what is said and how it is said. This ensures a more nuanced and accurate evaluation of interview performance.

3.4 Scoring

The AI Interviewer assigns a final score to each interview sample by integrating outputs from the modular analysis modules. The scoring formula weighs each component as follows:

- Text and audio sentiment (50%),
- Personality trait analysis (30%),
- Emotion detection (10%), and
- MFCC-based audio quality assessment (10%).

These weights were chosen informed by the relative explanatory power of each signal within interview-based communication as found in literature [41–43]. Sentiment, Personality and Emotion are interconnected but reveal influence in different circumstances. For example, sentiment refers to an individual's attitude towards the satisfaction of objective things or events, personality refers to an individual's consistent patterns of thought and feelings, and emotion refers to an individual's short-lived experience that can flactuate rapidly [43,44]. Following this, Sentiment was assigned the highest weight because lexical and paralinguistic sentiment remain the most direct indicators of how candidates formulate responses and convey evaluative stance. In other words, sentiment reflects positive and negative coverage of an individual's behaviour that affects managerial decision [43]. Personality traits contribute a substantial proportion of the score, as they provide broader behavioural context that helps interpret fluctuations in sentiment across an interview [42]. Emotion detection and MFCC derived audio quality measures were each given lower weights, since they offer complementary but secondary information about vocal expression and recording characteristics. This weighting strategy represents a pragmatic balance between interpretability, computational efficiency, and the need to capture diverse communicative cues without overemphsising any single modality. Importantly, the weights are treated as configurable parameters so HR teams can adjust them to match role-specific priorities, and the experiments reported in Section 4 evaluate the system under these default weights rather than claiming they are universally optimal.

This balanced composition ensures that both linguistic content and vocal delivery influence the final evaluation. Thereby offers a more holistic measure of candidate performance. The weighted approach allows HR teams to tailor evaluations to specific roles, ensuring alignment with organisational priorities and candidate suitability.

The final score is presented as PDF report alongside individual component breakdowns. This allows HR personnel or evaluators to trace how the score was derived and understand the contributing behavioural and vocal factors. This transparency enables HR teams to make data-driven decisions with confidence and ensures that the evaluation process is both fair and insightful for candidate selection.

3.5 User Interface and Reporting

The AI Interviewer system features a user-friendly Graphical User Interface (GUI), which enables non-technical users such as HR professionals to interact with the system efficiently. As shown in Fig. 2, users can upload interview audio files in WAV format alongside optional text transcripts, initiate multimodal analysis by clicking the Analyze button. This intuitive design ensures accessibility, allowing HR staff to seamlessly integrate the tool into their workflows, even without technical expertise, thus broadening its practical application in recruitment settings.

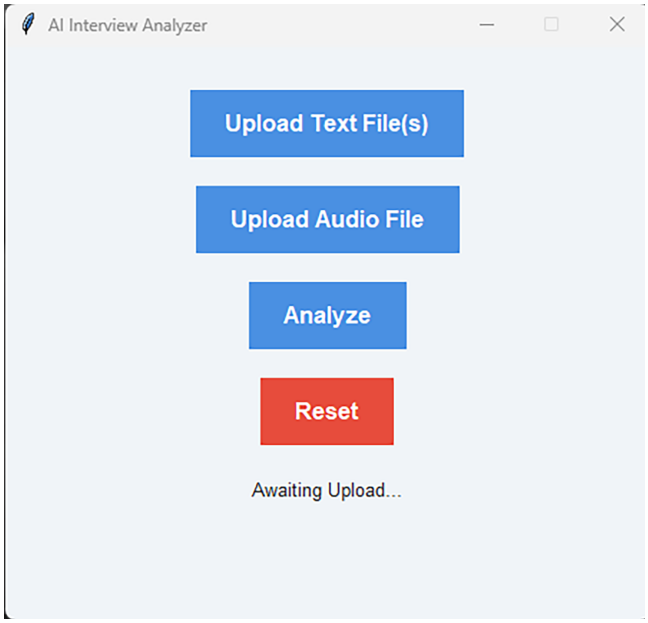

Figure 2: GUI of AI interviewer.

Upon completion of analysis, the system generates a comprehensive PDF report. As shown in Fig. 3, this report consolidates the outputs of all analytical modules, including sentiment scores (from VADER and transformer-assisted sentiment features where available), personality trait estimations, extracted keywords, full transcriptions, and audio features such as MFCCs. It also provides a detailed explanation of how the final score is computed. It should be noted, as mentioned in Section 3.4, weights applied to sentiment, personality and emotion in the calculation of final score could be treated as configuration parameter by HR team during the interview process. This is because the importance of sentiment, personality and emotion may vary depending on different job sectors. Overall, flexibility and clarity offered by our framework, foster trust and allow HR teams to confidently base decisions on the system's insights, ensuring evaluations are both reliable and actionable for recruitment purposes. The interface and reporting design prioritise usability, scalability, and clarity, ensuring the system can be adopted in real-world HR workflows while maintaining academic rigour in its evaluation process.

4 Evaluation of AI Interviewer

The evaluation phase rigorously assessed the AI Interviewer's ability to interpret and score interview responses using multimodal data: textual and acoustic inputs. It benchmarked the system against a unimodal baseline, where the baseline system comprises VADER and openSMILE. Evaluation was conducted using MELD to ensure a comprehensive analysis of sentiment detection capabilities [18]. This dataset choice reflects real-world conversational complexity. In addition, CMU-MOSI was used as an auxiliary benchmark to compare rule-based sentiment against transformer-based sentiment.

MP4 files were converted to WAV for MFCC extraction, and transcriptions were cleaned to remove errors and non-verbal markers such as laughter or sighs, ensuring data quality. A 5-fold cross-validation on MELD's training set refined the SVM emotion classifier, using stratified sampling to counter class imbalance. For sentiment classification improvements, additional probability-based transformer sentiment features were evaluated alongside VADER to strengthen performance on harder classes, including Negative sentiment.

AI Interview Analysis Report

Text Sentiment Analysis

N/A

Text Personality Analysis

N/A

Audio Sentiment Analysis

Overall Sentiment: POSITIVE (Score: 0.81)

Keywords: short flight

Audio Personality Analysis

Openness: High (0.81)

Conscientiousness: High (0.89)

Extraversion: High (0.85)

Agreeableness: High (0.97)

Neuroticism: Medium (0.35)

Audio Emotion Analysis

Detected Emotion: neutral (Score: 0.25)

Transcription

Flight? No, obviously it was a long time ago. The flight actually, because early in the year we actually went to Antigua in the Caribbean, which was the very first time we went, we actually had normal class because we went last year for two weeks just for two weeks we had normal class which is very stuffy because of all the legroom this time this time we actually went for an upgrade class for the business class so oh no this was this was their 2000 which was premium class which is okay and then this time, because it was only such a short flight, but we thought we needed it, we actually had the free seats by the door, which means we had extra legroom, which was actually very nice. The food actually on the flight was OK compared to other flights. This was to Majorca? Yeah, this was to Majorca. We flew with Air 2000. Yeah, it was from Gatwick South Terminal. How long? This one took under two hours, over two hours sometimes. It was OK. You drove around a bit? Yeah, drove around a bit. Had a quick look at the site and things. Our course Southampton, it's a two year course. First part of the course which is all done by the business students because you can have different sections like accounting and finance marketing. So basically I've actually went for marketing but for this one you can actually during the first year you can see what the other students have gone for doing and what bits they enjoyed and you can actually choose so you can choose whether you do marketing, finance and then you say I'd like to do this part. The quarter what you can do, you can do a work that will just just niche into business which means you can choose a free units. I think it's, well no it's actually yeah two units from each little section and Did you go to Southampton Institute. Is that part of the university? No, actually, Southampton Institute was actually Southampton Polytechnic, I think it was, or the older version of uni. So yeah, it was Southampton Polytechnic, then it became the Institute. And then, yeah, there's a sort of lot of there's a lot of there's a lot of there's a lot of rivalry comes spoke between the union Yeah Yeah, yeah, just just gonna just kind of pick it up. Yes Yeah, it should be should be very good and because it should be quite easy for me because I've been born in for the last Eight years free free prep school and five at Norton Abbey so I know what it's like to live away from home. So you'd be living in Southampton? Yeah, we're living in halls. It's not that far. It's basically hop on, hop off, go down to Southampton Central, hop on a train to find about a five minute train ride.

MFCC Features

MFCC Feature Count: 13

Mean MFCC Values (First 5): -213.72, 88.02, 29.12, 29.35, 11.49 ...

Final Score

Final Score Calculation:

- Sentiment (50%): 0.81 (Text: 0.50, Audio: 0.81)
- Personality (30%): 0.77 (Text: 0.50, Audio: 0.77)
- Emotion (10%): 0.25
- MFCC Quality (10%): 0.52

Formula: (0.5 * 0.81) + (0.3 * 0.77) + (0.1 * 0.25) + (0.1 * 0.52) = 0.71

Figure 3: Sample sentiment and personality scores (interview report).

Metrics included accuracy, macro F1-score, Cohen's Kappa, and class-specific precision, recall, and F1-scores for sentiment detection, offering a detailed performance overview. Confusion matrices and sentiment score distribution plots were also generated to visualise class-level behaviour and score separability.

4.1 Evaluation Metrics

To assess the performance and robustness of the AI Interviewer system, a comprehensive evaluation framework was established. The primary success indicator was the final composite score, which reflects an integrated view of the candidate's performance by combining sentiment polarity, emotion classification, personality inference, and audio quality indicators (such as MFCC consistency and vocal expressiveness). The goal was to ensure that the system produces coherent and interpretable scores that align with human expectations—for instance, assigning higher final scores to responses exhibiting positive emotional tone, clear articulation, and favourable personality traits (e.g., high agreeableness or conscientiousness). Furthermore, the system was evaluated for its resilience to missing data, ensuring that meaningful output is still produced when either the audio or text stream is incomplete. Sentiment classification was evaluated directly against labelled ground truth using standard classification metrics.

To support a fine-grained analysis, the following quantitative metrics were applied during evaluation:

- Sentiment Classification Metrics: The system's ability to accurately distinguish Positive, Neutral, and Negative sentiment was measured using Precision, Recall, Accuracy and Macro F1-score. These metrics were computed using standard definitions as shown below:

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN}, Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

and the F1-score for each class C (denotes the number of sentiment classes) is defined as:

$$F1_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

The macro F1-score is obtained by averaging across all classes:

$$\text{Macro} - \text{F1} = \frac{1}{C} \sum_{c=1}^{C} F1_c.$$

Results were drawn from both textual and audio sentiment predictions, verifying consistency across modalities. For MELD benchmarking, sentiment labels were evaluated at the utterance level against the dataset ground truth.

- Cohen's Kappa: To complement Accuracy and F1-based metrics, Cohen's Kappa (κ) was used to quantify agreement between the model's predicted sentiment labels and the ground-truth annotations. Unlike accuracy, Kappa adjusts for chance agreement, providing a more robust indicator of reliability in multi-class classification tasks. It is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o denotes the observed agreement between predictions and true labels, and p_e represents the expected agreement by chance. Higher values of κ indicate stronger consistency beyond random alignment. This metric is standard in sentiment and emotion classification research, particularly when class imbalance is present, and aligns with the reporting conventions used in MELD-based benchmarking.

- Emotion Detection Accuracy: Emotional states were evaluated using confusion matrices and category-level precision/recall, focusing on the system's ability to detect expressions of happiness, anger, sadness, and neutrality from acoustic features. These metrics were calculated using the predictions from the retrained Support Vector Machine (SVM) emotion classifier trained on MELD-labelled examples.
- Personality Trait Mapping Validity: The system generated inferred Big Five personality trait scores (openness, conscientiousness, extraversion, agreeableness, and neuroticism) from the sentiment analysis of candidate responses. These were cross-referenced with expected personality distributions from psycholinguistic literature to assess their construct validity, ensuring that positive sentiment correlated with traits such as extraversion and agreeableness, while negative sentiment aligned with higher neuroticism. Validation involved comparing the model's aggregated trait outputs with established linguistic–personality patterns, confirming that the inferred traits behaved in the expected directions across the evaluation samples.

Additionally, the quality of generated PDF reports was reviewed using a custom rubric focusing on clarity, structure, visual presentation, and the inclusion of actionable insights (e.g., interview improvement tips or emotional tone trends). The reports were rated highly for usability, confirming their utility as a communication bridge between algorithmic evaluations and human decision-making in recruitment scenarios.

4.2 Baseline Systems

To evaluate the performance of the AI Interviewer, a baseline model was implemented for comparative analysis. This baseline system utilises the VADER sentiment analysis tool, configured with the standard thresholds where compound scores above 0.05 are classified as positive, below −0.05 as negative, and values in between as neutral. In addition to VADER, the baseline incorporates an acoustic component extracted using

the openSMILE toolkit, which is employed to obtain MFCC features from the audio signal. A simple fusion strategy is applied whereby the mean MFCC value contributes a small additive adjustment to the VADER compound score, forming a lightweight VADER + openSMILE baseline. This configuration preserves the speed and simplicity of a text-driven model while allowing a minimal integration of vocal information for comparison with the full multimodal AI Interviewer.

Unlike the AI Interviewer, the baseline system uses pre-transcribed interview answers analysed by VADER, with only a very shallow acoustic adjustment based on the mean MFCC value. It does not exploit richer audio features such as pitch contours, energy dynamics, or full emotion classification, nor does it infer personality traits. This configuration therefore acts as a lightweight, weakly multimodal benchmark, capturing only coarse sentiment information compared with the AI Interviewer's full multimodal pipeline.

The baseline was chosen deliberately to reflect a common industry practice in lightweight automated assessments, where sentiment is derived solely from written content, an approach widely used in commercial text-based sentiment analysis tools and recruitment analytics platforms [45,46]. By comparing this standard approach, we demonstrate the added value of multimodal analysis in interview evaluation. The AI Interviewer's incorporation of audio cues, linguistic features, emotional classification, and psychological trait estimation provides a far more holistic candidate profile, which the baseline system cannot replicate. This contrast in complexity and dimensionality underscores the AI Interviewer's capacity to uncover subtleties that a text-only model may overlook.

4.3 Results

The AI Interviewer system achieved a classification accuracy of 0.64, macro F1-score of 0.61, and a Cohen's Kappa of 0.40 on the MELD test split, marking an improvement over the established baseline model (VADER + openSMILE), which reached an accuracy of 0.44, F1-score of 0.43, and Kappa of 0.16 (see Table 2). These results indicate that the proposed multimodal pipeline improves agreement with ground-truth sentiment labels and achieves stronger balanced performance across classes than a lightweight rule-based baseline augmented with shallow acoustic features. The results underline the system's effectiveness in multimodal fusion, combining text, audio, and emotion features for richer insights. This improvement is primarily driven by stronger separation of Neutral and Positive classes, while Negative sentiment remains inherently difficult due to restrained linguistic expression in interview contexts.

Table 2: Overall performance metrics.

Performance Metrics	AI Interview	VADER + openSMILE
Accuracy	0.64	0.44
Macro F1-Score	0.61	0.43
Cohen's Kappa	0.40	0.16

Class-wise performance provides further insights into how the model handles each sentiment (see Table 3):

Table 3: Class wise performance metrics.

	AI Interview			VADER + openSMILE		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Negative	0.61	0.49	0.55	0.42	0.50	0.46
Neutral	0.66	0.78	0.71	0.59	0.39	0.47
Positive	0.62	0.53	0.57	0.30	0.45	0.36

Although the Negative class yielded relatively high precision, the recall (0.4926) is comparable to the baseline (0.5037), indicating that both systems still miss a portion of subtly Negative samples. Neutral sentiment achieved the highest recall (0.7759), substantially higher than the baseline's (0.3875), aligning with the project's goal of better handling neutral tones, which are prevalent in professional and interview contexts where candidates often maintain composure. Positive sentiment recall (0.5323) also improved over the baseline (0.4540), suggesting better separation of Positive from Neutral and Negative cases on the MELD test split. Confusion matrices (Fig. 4) offer a visual comparison of class prediction distributions. Notably, the AI Interviewer correctly predicted 959 Neutral samples, in contrast to the baseline's 479, suggesting significant progress in moderate tone recognition, which is important for HR evaluations. However, the baseline shows slightly higher Negative recall (0.5037 vs. 0.4926), with 411 true Negative predictions vs. the AI Interviewer's 402, reveals a potential area for optimisation, especially in capturing subtle discontent or critique often masked in interviews.

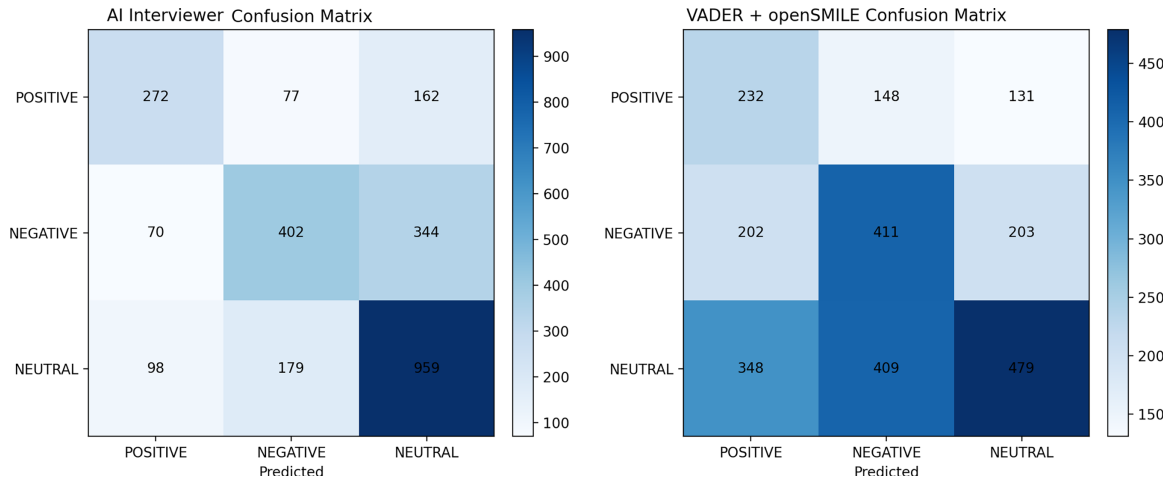

Figure 4: Comparison of confusion matrices.

The sentiment score distribution (Fig. 5) highlights differences in score calibration. The baseline (VADER + openSMILE) produces a strong concentration of scores around 0.0, reflecting the discrete behaviour of lexicon-based scoring and its tendency to assign many utterances near the neutral boundary. In contrast, the AI Interviewer yields a smoother and more continuously distributed score profile, indicating a more graded separation between sentiment intensities. This behaviour aligns with the confusion-matrix evidence: the baseline often pushes Neutral instances into Positive or Negative labels, whereas the AI Interviewer more frequently retains Neutral predictions, improving reliability for interview-style responses where moderate tone is common, enhancing its reliability for HR use.

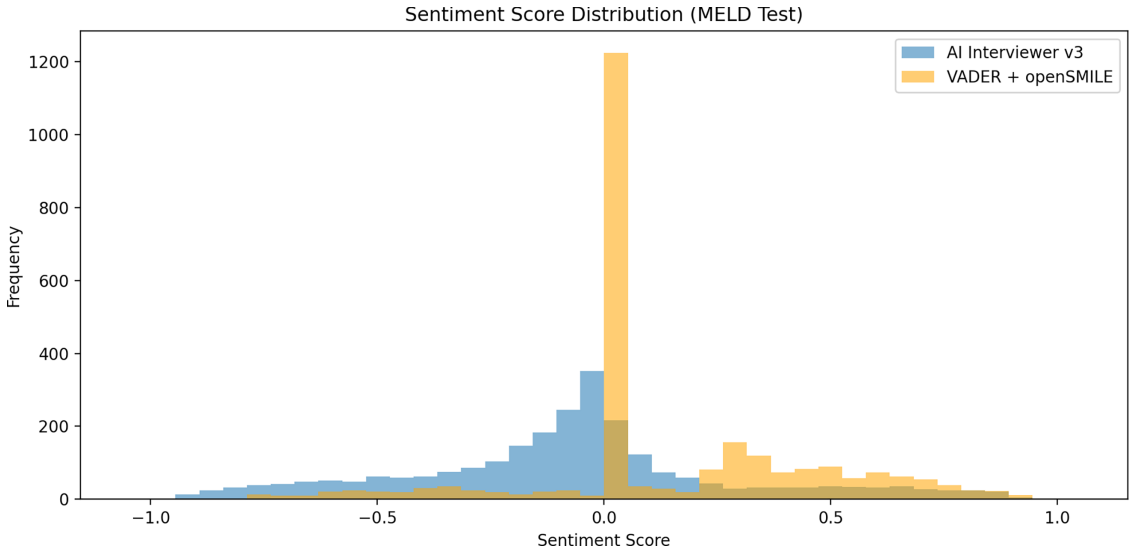

Figure 5: Sentiment score distribution (MELD Test).

4.4 Discussion

The evaluation results confirm that the AI Interviewer achieves a robust, multimodal framework for candidate interview assessment, surpassing unimodal methods and aligning with multimodal fusion research [26,47]. By integrating text and audio, it captures verbal and vocal nuances, offering a more comprehensive evaluation than unimodal systems that often miss critical emotional context. This holistic approach ensures a balanced assessment, considering both explicit content and implicit emotional cues, which is vital for understanding candidate suitability in HR contexts. The system's design reflects the growing trend in affective computing, where combining modalities enhances the accuracy of emotional analysis, providing a more reliable tool for recruitment evaluations.

Its high Neutral sentiment recall (0.78) improves over the baseline (0.39), suiting HR contexts with composed responses typical of professional settings [18]. This matters in interview settings, where responses are often controlled and non-polar. The confusion matrices reinforce this: the baseline misclassifies a large portion of Neutral samples into Positive or Negative, while the AI Interviewer substantially reduces this polarisation error, leading to more stable assessments for professionally phrased answers. Negative sentiment remains the most challenging class in terms of confusion with Neutral. While the AI Interviewer improves Negative precision (0.61 vs. 0.42), Negative recall is broadly comparable to the baseline (0.49 vs. 0.50), suggesting that subtle negativity expressed through restrained language or delivery can still be missed by both approaches (Table 3). Fig. 4 shows that many Negative instances are predicted as Neutral (e.g., 344 cases for the AI Interviewer), indicating that further work is needed to better separate mild critique or dissatisfaction from genuinely neutral responses. This pattern remains consistent with MELD's class distribution and conversational ambiguity, underscoring the need for stronger handling of nuanced negative cues in real-world interview scenarios. A plausible improvement path is to strengthen context modelling for borderline Negative cases while preserving Neutral robustness. One direction is to replace or augment rule-based sentiment features with transformer-derived probabilities (e.g., RoBERTa/DeBERTa-style contextual encoders) to better capture indirect negativity, hedging, and pragmatic cues common in formal interviews. In parallel, the fusion head could be extended with calibration techniques (e.g., temperature scaling) and cost-sensitive optimisation targeted at Negative–Neutral separation, ensuring that improvements in Negative detection do not reintroduce the baseline's tendency to polarise Neutral content.

Dataset rebalancing and weighted loss functions fell short, indicating limited generalisation for nuanced tones like indirect negativity often masked by neutral phrasing. This limitation suggests that the model struggles with contextual ambiguity, a common challenge in professional settings. The audio-text fusion, with a 0.4 audio weight rising to 0.5 for Negative cases with Neuroticism >0.7, may suppress subtle linguistic cues, prioritising vocal signals and risking misinterpretation of complex responses. Future iterations could incorporate attention mechanisms to dynamically balance modalities, ensuring both text and audio contribute proportionately to the final assessment, thus enhancing the system's sensitivity to layered emotional expressions.

VADER's lightweight structure ensures efficiency but misses sarcasm and irony, which candidates may use to express discontent. We tested DistilBERT, but did not use for real-time constraints [9]. This limitation can lead to overlooking critical emotional undertones, particularly in formal interviews. Future hybrid transformer approaches, like RoBERTa, could boost Negative recall by capturing deeper context [48], and LSTMs may enhance emotion detection by modelling vocal patterns over time [49]. These advancements would enable the system to better interpret subtle emotional cues, improving its applicability across diverse HR scenarios.

5 Conclusion and Recommendation for Future Work

In this paper, we present the design, implementation, and evaluation of AI Interviewer by applying a multimodal approach. The system presents a scalable, user-friendly solution with real-world applicability in candidate assessment, self-improvement, and corporate training. Its strengths include modularity, explainability, and cross-modal integration. Evaluation results show AI Interviewer surpassed the MELD baseline, demonstrating the viability of emotion-aware candidate evaluation. This is a significant achievement given the technical complexity involved in synchronising natural language processing (NLP), speech emotion recognition (SER), feature extraction, and GUI design into a unified workflow.

As AI continues to influence hiring processes, tools like the AI Interviewer, when ethically deployed, could offer consistent, bias-aware feedback, helping candidates refine soft skills and providing organisations with richer insights into interpersonal fit. Ethical considerations were paramount throughout the design and development of the AI Interviewer. For example, the datasets we used in our framework are widely used in sentiment and emotion research and were selected due to their detailed emotional annotations and lack of personally identifiable information (PII), ensuring privacy compliance. In order to mitigate bias, we have taken proactive steps by balancing sentiment classes during training and introducing personality-based adjustment logic. For example, negative sentiment predictions were weighted against indicators such as high Neuroticism (above 0.7), preventing unfair penalisation of emotionally expressive candidates.

Despite its achievements, the system exhibited limitations. The recall for the Negative sentiment class remained comparable to the baseline, indicating that both systems still miss a portion of subtly Negative samples in formal interview-style responses. This opens a future research direction for AI Interviewer, where a hybrid sentiment analysis architecture could be adopted that combines rule-based logic (e.g., VADER) with transformer-based contextual deep learning models, such as RoBERTa [48] or DeBERTa [50], both of which have shown superior contextual accuracy in emotion-rich environments. These models can help detect negativity that may be expressed in subtle, indirect, or linguistically ambiguous ways, especially in formal settings like interviews.

Additionally, further work could fine-tune the SVM classifier used for emotion recognition or replace it entirely with deep learning architectures, particularly models like wav2vec2 [51] or ECAPA-TDNN [52], which have demonstrated high accuracy in speech emotion recognition using self-supervised learning from raw audio waveforms.

There is also notable potential in advancing the personality trait inference, which currently derives OCEAN dimensions from average sentiment scores. Future systems could integrate psycholinguistic models based on LIWC (Linguistic Inquiry and Word Count) [53] or OpenPersonality datasets, enabling finer-grained interpretation of features such as function word usage, abstraction level, and interpersonal reference patterns [54]. This would enable the AI Interviewer to provide more accurate, multi-dimensional insights into a candidate's interpersonal style and behavioural tendencies.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Daniil Kamakaev; methodology, Daniil Kamakaev, and Khaled Mahbub; software, Daniil Kamakaev; validation, Daniil Kamakaev; formal analysis, Daniil Kamakaev, and Khaled Mahbub; investigation, Daniil Kamakaev; resources, Daniil Kamakaev; data curation, Daniil Kamakaev; writing, Daniil Kamakaev, and Khaled Mahbub; writing—review and editing, Khaled Mahbub; visualization, Daniil Kamakaev, and Khaled Mahbub; supervision, Khaled Mahbub; project administration, Khaled Mahbub. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available as described below: Multimodal EmotionLines Dataset (MELD), https://affective-meld.github.io/. The Interactive Emotional Dyadic Motion Capture (IEMOCAP) Database, https://sail.usc.edu/iemocap/. CMU-MOSI Dataset, http://multicomp.cs.cmu.edu/resources/cmu-mosi-dataset/.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Schmidt FL, Hunter JE. The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol Bull. 1998;124(2):262–74. doi:10.1037/0033-2909.124.2.262.
2. Dastin J. Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. 2018 [cited 2025 Jan 1]. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
3. Hewage A. The applicability of artificial intelligence in candidate interviews in the recruitment process. J Manag Stud Dev. 2023;2(2):174–97. doi:10.56741/jmsd.v2i02.388.
4. Chen Z. Collaboration among recruiters and artificial intelligence: removing human prejudices in employment. Cogn Technol Work. 2023;25(1):135–49. doi:10.1007/s10111-022-00716-0.
5. Mori M, Sassetti S, Cavaliere V, Bonti M. A systematic literature review on artificial intelligence in recruiting and selection: a matter of ethics. Pers Rev. 2025;54(3):854–78. doi:10.1108/pr-03-2023-0257.
6. Dadaboyev SMU, Abdullayeva J, Abbosova N, Suleymenova A, Mamadjanova K. Role of artificial intelligence in employee recruitment: systematic review and future research directions. Discov Glob Soc. 2025;3(1):99. doi:10.1007/s44282-025-00246-w.
7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. 2018. doi:10.48550/arXiv.1810.04805.
8. Sanh V, Debut L, Chaumond J, Wolf T. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108. 2019. doi:10.48550/arXiv.1910.01108.
9. Hutto CJ, Gilbert E. VADER: a parsimonious rule-based model for sentiment analysis of social media text. In: Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media; 2014 Jun 1–4; Ann Arbor, MI, USA; 2014. p. 216–25.

10. Schuller B, Steidl S, Batliner A, Vinciarelli A, Scherer K, Ringeval F, et al. The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, and adaptation. In: Proceedings of the INTERSPEECH 2013; 2013 Aug 25–29; Lyon, France. p. 3244–8. doi:10.21437/Interspeech.2013-56.
11. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in openSMILE, the munich open-source multimodal feature extractor. In: Proceedings of the 21st ACM International Conference on Multimedia; 2013 Oct 21–25; Barcelona, Spain. p. 835–8. doi:10.1145/2502081.2502224.
12. Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust Speech Signal Process. 1980;28(4):357–66. doi:10.1109/tassp.1980.1163420.
13. HireVue. HireVue: AI-powered talent assessment solutions. 2023 [cited 2025 Jan 1]. Available from: https://www.hirevue.com/.
14. Jamal A. Mya systems—"using conversational AI to solve talent acquisition challenges". 16 Apr 2022 [cited 2025 Jan 1]. Available from: https://d3.harvard.edu/platform-digit/submission/mya-systems/.
15. Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency; 2018 Feb 23–24; New York, NY, USA. p. 1–15. doi:10.1145/3287560.3287596.
16. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1(5):206–15. doi:10.1038/s42256-019-0048-x.
17. Busso C, Bulut M, Lee CC, Kazemzadeh A, Mower E, Kim S, et al. IEMOCAP: interactive emotional dyadic motion capture database. Lang Resour Eval. 2008;42(4):335–59. doi:10.1007/s10579-008-9076-6.
18. Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: a multimodal multi-party dataset for emotion recognition in conversations. arXiv:1810.02508. 2019. doi:10.48550/arXiv.1810.02508.
19. Grootendorst M. KeyBERT: minimal keyword extraction with BERT. GitHub Repository. 2020 [cited 2025 Jan 1]. Available from: https://github.com/MaartenGr/KeyBERT.
20. McFee B, Raffel C, Liang D, Ellis D, McVicar M, Battenberg E, et al. Librosa: audio and music signal analysis in Python. In: Proceedings of the 14th Python in Science Conference; 2015 Jul 6–12; Austin, TX, USA. p. 18–24. doi:10.25080/majora-7b98e3ed-003.
21. Zadeh A, Zellers R, Pincus E, Morency L. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv:1606.06259. 2016. doi:10.48550/arXiv.1606.06259.
22. Liu B. Sentiment analysis and opinion mining. In: Synthesis lectures on human language technologies. Cham, Switzerland: Springer; 2012. p. 1–167. doi:10.1007/978-3-031-02145-9.
23. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD. Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language ProcessingSentiment Analysis and Opinion Mining. Washington, DC, USA; 2013. p. 1631–42.
24. Antoniou N, Katsamanis A, Giannakopoulos T, Narayanan S. Designing and evaluating speech emotion recognition systems: a reality check case study with IEMOCAP. arXiv:2304.00860. 2020. doi:10.48550/arxiv.2304.00860.
25. Ahmad Al Hindawi N, Shahin I, Bou Nassif A. The exploitation of multiple feature extraction techniques for speaker identification in emotional states under disguised voices. In: Proceedings of the 2021 14th International Conference on Developments in eSystems Engineering (DeSE); 2021 Dec 7–10; Sharjah, United Arab Emirates. p. 269–73. doi:10.1109/dese54285.2021.9719357.
26. Poria S, Cambria E, Bajpai R, Hussain A. A review of affective computing: from unimodal analysis to multimodal fusion. Inf Fusion. 2017;37:98–125. doi:10.1016/j.inffus.2017.02.003.
27. Wöllmer M, Weninger F, Knaup T, Schuller B, Sun C, Sagae K, et al. YouTube movie reviews: sentiment analysis in an audio-visual context. IEEE Intell Syst. 2013;28(3):46–53. doi:10.1109/MIS.2013.34.
28. Mittal T, Bhattacharya U, Chandra R, Bera A, Manocha D. M3ER: multiplicative multimodal emotion recognition using facial, textual, and speech cues. arXiv:1911.05659. 2019. doi:10.48550/arxiv.1911.05659.

29. Ganesh Kumar P, Arul Antran Vijay S, Jothi Prakash V, Paul A, Nayyar A. A context-sensitive multi-tier deep learning framework for multimodal sentiment analysis. Multimed Tools Appl. 2024;83(18):54249–78. doi:10.1007/s11042-023-17601-1.
30. Jothi Prakash V, Arul Antran Vijay S. A novel approach to cross-linguistic transfer learning for hope speech detection in Tamil and Malayalam. Comput Speech Lang. 2026;96:101870. doi:10.1016/j.csl.2025.101870.
31. Goldberg LR. An alternative "description of personality": the Big-Five factor structure. J Pers Soc Psychol. 1990;59(6):1216–29. doi:10.1037/0022-3514.59.6.1216.
32. Zhang X, Wang MC, He L, Jie L, Deng J. The development and psychometric evaluation of the Chinese Big Five Personality Inventory-15. PLoS One. 2019;14(8):e0221621. doi:10.1371/journal.pone.0221621.
33. Xing W, Liu C, Zhang K, Peng Y, Li X, Li K, et al. The association between Big Five personality traits and social anxiety symptoms in Chinese college students: a network analysis. Preprint. 2024. doi:10.21203/rs.3.rs-3933883/v1.
34. Zell E, Lesick TL. Big five personality traits and performance: a quantitative synthesis of 50+ meta-analyses. J Pers. 2022;90(4):559–73. doi:10.1111/jopy.12683.
35. Barrick MR, Mount MK. The big five personality dimensions and job performance: a meta-analysis. Pers Psychol. 1991;44(1):1–26. doi:10.1111/j.1744-6570.1991.tb00688.x.
36. Rothmann S, Coetzer EP. The big five personality dimensions and job performance. SA J Ind Psychol. 2003;29(1):68–74. doi:10.4102/sajip.v29i1.88.
37. Abdul ZK, Al-Talabani AK. Mel frequency cepstral coefficient and its applications: a review. IEEE Access. 2022;10:122136–58. doi:10.1109/access.2022.3223444.
38. Deruty E. Intuitive understanding of MFCCs. 16 Sep 2022 [cited 2026 Jan 1]. Available from: https://medium.com/p/836d36a1f779.
39. Athina B. Audio signal feature extraction for analysis. 4 Mar 2020 [cited 2026 Jan 1]. Available from: https://athina-b.medium.com/audio-signal-feature-extraction-for-analysis-507861717dc1.
40. Ekman P. An argument for basic emotions. Cogn Emot. 1992;6(3–4):169–200. doi:10.1080/02699939208411068.
41. Choudhury P, Wang D, Carlson N, Khanna T. Machine learning approaches to facial and text analysis: discovering CEO oral communication styles. SSRN J. 2019. doi:10.2139/ssrn.3392448.
42. Wehner C, de Grip A, Pfeifer H. Do recruiters select workers with different personality traits for different tasks? A discrete choice experiment. Labour Econ. 2022;78(2):102186. doi:10.1016/j.labeco.2022.102186.
43. Ali Akber M, Ferdousi T, Ahmed R, Asfara R, Rab R, Zakia U. Personality and emotion—a comprehensive analysis using contextual text embeddings. Nat Lang Process J. 2024;9(7):100105. doi:10.1016/j.nlp.2024.100105.
44. Ma Z, Jia W, Zhou Y, Xu B, Liu Z, Wu Z. Personality enhanced emotion generation modeling for dialogue systems. Cogn Comput. 2024;16(1):293–304. doi:10.1007/s12559-023-10204-w.
45. Cambria E, Das D, Bandyopadhyay S, Feraco A, editors. Affective computing and sentiment analysis. In: A practical guide to sentiment analysis. Cham, Switzerland: Springer International Publishing; 2017. p. 1–10. doi:10.1007/978-3-319-55394-8_1.
46. Ribeiro FN, Araújo M, Gonçalves P, André Gonçalves M, Benevenuto F. SentiBench—a benchmark comparison of state-of-the-practice sentiment analysis methods. EPJ Data Sci. 2016;5(1):23. doi:10.1140/epjds/s13688-016-0085-1.
47. Zadeh A, Chen M, Poria S, Cambria E, Morency L-P. Tensor fusion network for multimodal sentiment analysis. arXiv:1707.07250. 2017. doi:10.48550/arxiv.1707.07250.
48. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019. doi:10.48550/arXiv.1907.11692.
49. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. DialogueRNN: an attentive RNN for emotion detection in conversations. arXiv:1811.00405. 2019. doi:10.48550/arXiv.1811.00405.
50. He P, Liu X, Gao J, Chen W. DeBERTa: decoding-enhanced BERT with disentangled attention. arXiv:2006.03654. 2020. doi:10.48550/arXiv.2006.03654.
51. Baevski A, Zhou H, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. arXiv:2006.11477. 2020. doi:10.48550/arxiv.2006.11477.
52. Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. arXiv:2005.07143. 2020. doi:10.48550/arXiv.2005.07143.

53. Tausczik YR, Pennebaker JW. The psychological meaning of words: LIWC and computerized text analysis methods. J Lang Soc Psychol. 2010;29(1):24–54. doi:10.1177/0261927x09351676.
54. Schwartz HA, Eichstaedt JC, Kern ML, Dziurzynski L, Ramones SM, Agrawal M, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS One. 2013;8(9):e73791. doi:10.1371/journal.pone.0073791.